# The internal validity of web-based studies

PHILIPPE F. LACHEREZ
*School of Psychology, The University of Queensland*

ABSTRACT: Honing and Ladinig (2008) make the assertion that while the internal validity of web-based studies may be reduced, this is offset by an increase in external validity possible when experimenters can sample a wider range of participants and experimental settings. In this paper, the issue of internal validity is more closely examined, and it is agued that there is no necessary reason why internal validity of a web-based study should be worse than that of a lab-based one. Errors of measurement or inconsistencies of manipulation will typically balance across conditions of the experiment, and thus need not necessarily threaten the validity of a study's findings.

Submitted 2008 August 2; accepted 2008 August 3.

KEYWORDS: *internal validity, reliability, web-based studies*

IN the discussion by Honing & Ladinig (2008, see also Kendall, 2008; Honing & Reips, 2008) the authors consider previous criticisms of web-based research concerned with the purported lack of experimental control involved in such research. The authors make the laudable point that such lack of control is offset by the increase in generalizability and ecological validity brought about by the greater range of apparatus and settings possible with presentation over the web. They conclude that internal validity may be threatened, but that external validity is increased, when a greater variety of settings is possible. The purpose of this commentary is to take up the question of whether internal validity is necessarily threatened when experiments lack laboratory control, and therefore whether such lack of control should necessarily challenge the interpretability of the findings of web-based studies.

The *internal validity* of a study is the extent to which it is possible to say that the treatment effects observed in an experiment are due to the treatment (intended manipulation) and not some other factor introduced by the manipulation (Campbell & Stanley, 1963). Threats to internal validity are variables (typically known as *confounds*) that vary systematically with the levels of the manipulation, thus producing significant effects, which lead the researcher to conclude their manipulation has been effective, while in reality another unobserved factor has been responsible for the outcome. Such confounds are a threat to the research community, as it is possible that researchers will mistakenly believe that a variable has an impact when in fact it does not.

For a variable to threaten internal validity of a study, the variable must vary systematically with the manipulation (for instance a confound that affects all participants in one condition differently to those in another condition). Nuisance variables which vary randomly within conditions produce greater error variance for analyses, and make it less likely the experimenter will observe a significant finding, by creating a loss of power to detect differences between conditions. An issue such as differences in headphones between participants, for instance, cannot itself vary systematically across conditions (i.e., to produce a false significant result) unless it is *systematically* the case that participants in one condition of the experiment consistently wore one kind of headphones, while the participants in the other condition wore a second kind. If the headphones themselves were simply randomly different across all participants, this could not produce a false significant result, but would merely create error within each condition. These would be more likely to create non-significant results, if such artifacts were inflating error, because they make the differences between subjects in each condition larger, and therefore the statistic observed (which will always be some ratio of differences between treatments to differences within treatments) smaller. The same could be said of any artifact of lack of experimental control in such studies. Unless the supposed artifact (person differences, headphones, room acoustics, computer speed, etc.) clearly varied systematically between conditions, there is no reason why it should influence the interpretability of a significant difference if one were observed.

It is possible that there might be some circumstances in which an artifact of sampling, or of instrumentation, could vary systematically between conditions, but off hand it is hard to imagine such a circumstance. As Honing and Ladinig (2008) point out, participants may be able to guess the hypotheses of studies that do not disguise their intention, but such a confound is well understood by lab researchers and so it is imagined that whatever procedures the researchers currently use to prevent hypothesis guessing, can also be used over the web. It is simply beholden on the researcher to consider

the same kinds of potential confounding variables in a web-based study that they would consider in a laboratory-based one.

It is true that not being able to control sources of error variance does reduce the power of web-based experiments. Imagine a clearly defined experimental task, in which participants are presented with clearly contrasting sounds. In a laboratory, it may be the case that an experimenter would observe a strong significant level of discrimination in all participants. Outside the laboratory, with other contaminating influences of ambient noise, differences in apparatus, and so on, one or more participants may fail to observe the differences, producing a weaker result. However, a benefit of web-based experiments is that in general they allow for greater subject numbers because it is possible to recruit participants from a wider area (lab-based studies are restricted to those participants who are physically able to be present) and participants may generally be happier to volunteer if they are able to complete the experiments in the comfort of their own home. The larger sample sizes possible with web experiments may help to remedy some problems of weaker power due to less control, because larger N will produce greater statistical power for the same effect size. Moreover the larger sample size also helps to more effectively balance these sources of error variance across conditions.

In regard to the issue of *reliability* raised by Kendall (2008), it is true, as Honing and Reips (2008) point out, that a construct must be measured reliably for it to be measured accurately (*construct validity*). Again, however, any inaccuracy of measurement would have to be consistent *within* conditions of the experiment, and differential *across* conditions, for this to create false significant findings in a study. Measurement error within conditions cannot itself confound a study, but merely reduce the power of analyses to find significant effects. It is possible that an artifact may lead to an experimenter being mistaken about *what* was being measured (for instance response times that are affected by differences in download speed in the different conditions of the experiment), but this is not an issue of reliability. Rather, such issues of interpretation only manifest when the observer has *reliably* measured the wrong thing (Mitchell & Jolley, 2007, Babbie, 2007). These issues of construct validity should be dealt with in the usual ways (by using manipulation checks, convergent and divergent measures, etc., see Mitchell & Jolley, 2007, for a discussion), and are not unique to web-based studies.

In short, because increases in error arising from the lack of control in a web experiment can usually be expected to vary randomly within conditions of the experiment (unless a true confound is present owing to an unintended manipulation), there is no greater danger of false significant effects in a web-based study than in a lab-based one. There is potentially less power for detecting effects when data are noisier or are contaminated by factors outside the experimenter's control. As such, experimenters wishing to find strongly significant findings in controlled settings should be aware of this potential reduction in power. In either case, it is important to note that it is not the validity of the findings that is threatened, but merely the power of the analyses to detect differences. Thus, research using web-based protocols should not be dismissed merely because there is a greater potential for random errors within conditions. Rather, it is only when there is reason to believe that differences in such artifacts may occur *between* conditions of the experiment that a critic need be concerned about the validity of the findings.

## REFERENCES

Babbie, E. R. (2007). *The practice of social research.* Belmont, CA: Thomson Wadsworth.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, Rand McNally.

Honing, H., & Ladinig, O. (2008). The potential of the Internet for music perception research: A comment on lab-based versus Web-based studies. *Empirical Musicology Review,* Vol. 3, No. 1*,* 4-7.

Honing, H., & Reips, U. (2008). Web-based versus lab-based studies: A response to Kendall (2008). *Empirical Musicology Review,* Vol. 3, No. 2*,* 73-77.

Kendall, R. A. (2008). Commentary on "The potential of the Internet for music perception research: A comment on lab-based versus Web-based studies" by Honing & Ladinig. *Empirical Musicology Review,* Vol. 3, No. 1*,* 8 – 10.

Mitchell, M. L., & Jolley, J. M. (2007). *Research design explained*. Belmont, CA: Thomson Wadsworth.